Lecture 20

Desriptive Statistics

Manju M. Johny

STAT 330 - Iowa State University

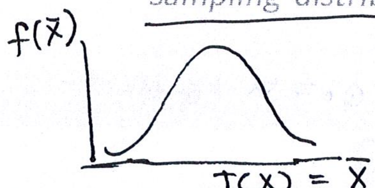
Statistics

Statistics

Definition: Statistics

A *statistic*, $T(X_1, \ldots, X_n)$ is a function of random variables.

- Start with taking a *simple random sample (SRS)* of size n from some population/distribution.
  $X_1, \ldots, X_n \overset{iid}{\sim} f_X(x)$
- We can then obtain *statistics* based on $X_1, \ldots, X_n$ $= T(X_1 \ldots X_n)$
- Since a statistic is a function $T(\cdot)$ of random variables, the statistic is also a random variable.
- Thus, the statistic will have its own distribution called the *sampling distribution of the statistic* (more on this later!)



Statistics Cont.

Definition: Observed Statistics

The *observed statistics*, $T(x_1, \ldots, x_n)$ is the statistic function with observed values plugged in.

- *Descriptive statistics:* Describing what our sample data looks like (graphically or numerically)
- *Inferential statistics:* Use the statistic to infer/learn about the "true" distribution, $f_X(x)$, that generated the data.

Note:

- Use small letters (x, $\bar{x}$, s^2, etc) to represent observations and observed statistics.
- Use capital letters (X, $\bar{X}$, S^2, etc) to represent random variables.

Mean and Variance

Sample Mean and Variance

Let $X_1, \ldots, X_n \overset{iid}{\sim} f_X(x)$ where $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$

- *Sample mean* is defined as $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - → estimates the population mean μ.
- *Sample variance* is defined as $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$
  - → estimates the population variance σ^2
  - → an estimate of the $Var(X) = E[(X - E(X))^2]$ can be found as $\frac{1}{n}\sum_{i=1}^{n}(X_i - (\bar{X}))^2$
  - → typically, n in the above denominator is replaced with $n - 1$ to get S^2 (more on this later)
- *Sample standard deviation* is $S = \sqrt{S^2}$

Note: The quantities above are R.V's since they are functions of R.V's $X_1, \ldots, X_n$.

Observed Sample Mean and Variance

- To obtain the *observed sample mean* and *observed sample variance*, plug in observed data values $(x_1, \ldots, x_n)$ into sample mean and variance formulas

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

observed sample mean

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

observed sample variance

$$s = \sqrt{s^2}$$

observed sample standard dev.

Note: The quantities above are not random variables since you have plugged in data values. They are values such as 2.4, 100, *etc.*

Quantiles

Quantiles

Definition: Quantiles

The q^{th} *quantile* of a distribution, $f_X(x)$, is a value x such that $P(X < x) \leq q$ and $P(X > x) \leq 1 - q$.

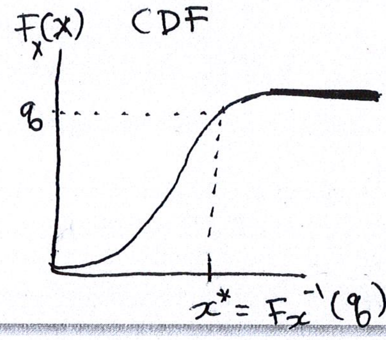
This is also called the $100 \cdot q^{th}$ *percentile.*

$Q_1 = 0.25^{th}$ quantile, $Q_2 = 0.5^{th}$ quantile (median), and $Q_3 = 0.75^{th}$ quantile

Definition: Quantile Function

The *quantile function* is defined as:

$F_X^{-1}(q) = min\{x : F_X(x) \geq q\}$



Median

50 % of my X's are less than median

$X_{(k)}$ is an "order" statistic

$X_{(1)} = \min$

$X_{(n)} = \max$

$X_{(5)} = 5^{th}$ ordered value

The *median* is the 0.5^{th} quantile (or 50^{th} percentile)
→ can be written as $F_X^{-1}(0.5)$
The *sample median* is calculated by:

1. Order sampled values in increasing order: : $X_{(1)}, \ldots, X_{(n)}$
   - If n is odd, take the middle value
     → median $= X_{\lceil \frac{n}{2} \rceil}$ ← $X_{\lceil n/2 \rceil}$

     2.2 (4.1) 7.3
     med = 4.1
   - If n is even, average the two middle values
     → median $= \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}$ ← $\frac{X_{(n/2)} + X_{(n/2+1)}}{2}$

     2.2 (3 4) 5.7
     median = 3.5

Note: Since the above values are functions of R.V's, they are R.Vs. Obtain the *observed sample median* by plugging in the observed values $(x_1, \ldots, x_n)$ from data.

Q_1 and Q_3

Other sample quantiles we are typically interested in are

- $Q_1 = 0.25^{th}$ quantile
- $Q_3 = 0.75^{th}$ quantile

Many ways to calculate quantiles. Our method for a general q^{th} sample quantile is ...

1. Compute $(n+1) \cdot q$
   - If this value is an integer, use $(n+1)q^{th}$ ordered value
   - Else, use the average of the 2 surrounding values

Example

Example 1: A sample $X_1, \ldots, X_n \overset{iid}{\sim} f_X(x)$ was taken where $X_i =$ CPU time for a randomly chosen task. The ordered observed values are $15, 34, 35, 36, 43, 48, 49, 62, 70, 82$ (secs)

sample data

The observed...

- sample mean : $\bar{x} = \dfrac{15 + 34 + \cdots + 82}{10} = 47.4$
- sample variance : $s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \dfrac{(15-47.4)^2 + \cdots + (82-47.4)^2}{9}$ $= 384.04$
- sample std. dev : $s = \sqrt{s^2} = \sqrt{384.04}$
- sample median : $n = 10 \rightarrow$ even (pick out 2 middle values)

$$\text{med} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{43 + 48}{2} = 45.5$$

Example Cont.

- sample Q_1 : $(n+1)q = (10+1)(0.25) = 2.75$ ← find $x_{(2)}$ & $x_{(3)}$, & average them

$$Q_1 = \frac{x_{(2)} + x_{(3)}}{2} = \frac{34 + 35}{2} = 34.5$$

- Sample Q_3 : $(n+1)q = (10+1)(0.75) = 8.25$ ← average the 8th & 9th ordered values $x_{(8)}$, $x_{(9)}$

$$Q_3 = \frac{x_{(8)} + x_{(9)}}{2} = \frac{62 + 70}{2} = 66$$

Right now, we're only using these statistics to describe the sample of CPU speeds.

- sample mean and median (Q_2) tell us "typical" values
- sample variance tells us how "spread out" / how variable the data are
- Q_1 and Q_3 "rank" where values fall in our sample

Mode, Range, IQR

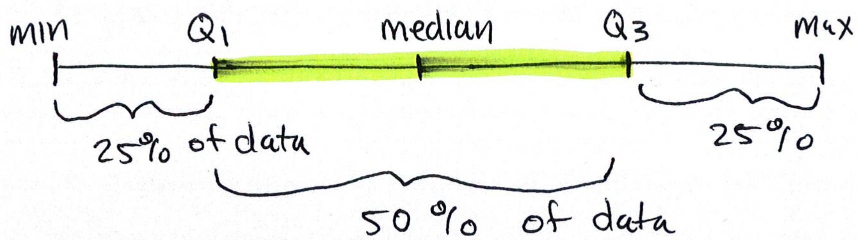
Mode, Range, and IQR

Other common descriptive statistics to describe the data:

- *Mode:* The most frequent value in our sample. Can have multiple modes in data set
- *Range:* Max - Min = $X_{(n)} - X_{(1)}$

  → describes the "total" variability of the data
- *Interquatrile Range (IQR):* $Q_3 - Q_1$

  → describes the variability of the middle 50% of data



Robust Statistics

- With all the different options for statistics, how do we choose which ones to use?

  → It depends on your data set
- Statistics that are not affected by extreme values are called *robust statistics*

Robust	Not Robust
median, IQR	mean, std. dev, range

Example 2:

Imagine Keanu Reeves moves into your nighborhood.

statistic	Pre-Keanu	Post-Keanu	Robust?
mean	$40K	way bigger	No
median	$ 40K	same or slightly bigger	Yes
standard dev	$ 10K	way bigger	No
IQR	$ 25K	same or slightly bigger	Yes